CROWDSOURCING BASED SUBJECTIVE QUALITY ASSESSMENT OF ADAPTIVE VIDEO STREAMING

M. Shahid , J. Søgaard , J. Pokhrel , K. Brunnström , K. Wang , S. Tavakoli , and N. Gracia

ABSTRACT

In order to cater for user's quality of experience (QoE) requirements, HTTP adaptive streaming (HAS) based solutions of video services have become popular recently. User QoE feedback can be instrumental in improving the capabilities of such services. Perceptual quality experiments that involve humans are considered to be the most valid method of the assessment of QoE. Besides lab-based subjective experiments, crowdsourcing based subjective assessment of video quality is gaining popularity as an alternative method. This paper presents insights into a study that investigates perceptual preferences of various adaptive video streaming scenarios through crowdsourcing based subjective quality assessment.

Index Terms— Adaptive streaming, Subjective, Video quality assessment, Crowdsourcing, Buffering

1. INTRODUCTION

Interest in quality of experience (QoE) of video services is growing due to increasing usage of videos over networks, such as the portion of video data in mobile networks is expected to exceed 67% by 2018 [1]. Hypertext Transfer Protocol (HTTP) based video streaming has been greatly adopted to avoid network distortions such packet-loss. Subjective experiments are considered to be the most valid methodology to assess the QoE. Subjective experiments are typically conducted in a controlled laboratory environment. Objective or computer software assisted methods have been largely seen as an alternative approach, to get around the complications involved in the lab-based subjective experiments. However, the objective methods even with state-of-the-art performance are generally considered far from universal acceptance. Crowdsourcing based subjective experiments have gained attention to replace needs of lab-based tests and these experiments offer promising correlation with the later [2]. This methodology mainly involves collecting subjective assessment of quality through ubiquitous streaming via the Internet. This enables the investigator to receive opinion from a vast variety of subjects; in a time-flexible, test-data size scalable, and swift manner.

This paper includes an insight into a crowdsourcing based subjective perceptual preference of various adaptation scenarios investigated earlier in [3] and additional buffering scenarios. In the following, we present the related details on our experiments and the obtained results thereof.

2. TEST BACKGROUND

The videos for our subjective test are originally from the subjective lab experiment detailed in [3]. The original videos were all in 1280x720 resolution with a frame rate of 24 and encoded using the high profile for H.264/AVC at 4 different bitrates: 600 kbps, 1 Mbps, 3 Mbps, and 5 Mbps. Seven different sources were used; three sources were taken from entertainment movies and the rest was content from: a soccer match, a sports documentary, a newscast, and a concert. The subjective lab experiment was carried out at Acreo lab in a test room compliant with the ITU-R BT.500 [4]. Several adaptation scenarios for the videos were produced in the original experiment, such as going from a high to a low bitrate in a stepwise manner. In our subjective experiment we used the following scenarios from the original experiment: Gradual Decreasing (GD), Rapid Decreasing (RD), constant 600 kbps (N600), constant 1 Mbps (N1), constant 3 Mbps (N3), and constant 5 Mbps (N5). Additionally, we introduced new buffering scenarios to test the quality perception in relation to the aforementioned scenarios. The buffering scenarios include: 1 Freezing event lasting for 2 seconds in the constant 3 Mbps video (1F3M), 2 Freezing events lasting for 1 second each in the constant 3 Mbps video (2F3M), and 1 Freezing event lasting for 2 seconds in the constant 1 Mbps video (1F1M). In total 9 different scenarios were used, resulting in a total of 63 stimuli.

Crowdsourcing experiments should be as simple as possible for the subject, therefore we chose to follow the Paired Comparison (PC) methodology [5]. We used the optimized square design [6] based on our assumptions of the quality levels to get reliable measurements and reduce the number of pairings. Using this method, our test set consisted of a total of 126 pairings. These pairing were divided into 14 tasks with 3 videos from 3 different contents, i.e., 9 videos for each task. We used screentests [7] prior to the subjective test to

filter out potential malicious workers. In total, 215 workers participated in the experiment where almost one quarter of the participants were from the US while the rest were mainly from European countries.

3. RESULTS

To analyze the results, we applied the Bradley-Terry (BT) model to obtain quality scores and corresponding confidence intervals from the preference matrices of PC data [6]. The subjects were filtered by excluding workers with too many unlikely preferences. We define an unlikely preference as a preference where the corresponding probability in the BT model is lower than a threshold θ. In our test, we allowed only 2 out of 9 unlikely preferences and therefore we set $\theta = 0.25$. With this approach 6 workers were excluded from the final results. In order to validate the results obtained from the crowdsourcing experiment, we compared the mean opinion scores (MOS) obtained from the lab experiment to the crowdsourcing experiment as shown in Fig. 1. The results show that the opinion scores obtained from both the experiments are strongly correlated, though not to the degree one would expect of a repetition of the lab test. This can be due to the differences in the test setup, such as evaluation method, viewing environment and the introduction of new distortions. Our experiment verify the results from earlier studies, e.g., [8], that buffering events has a high impact on the QoE. Due to this, users generally prefer viewing videos at lower bitrates than experiencing buffering events in videos at higher bitrates.

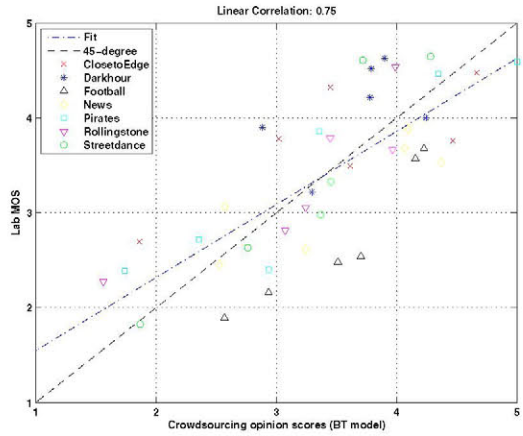

Fig. 1. Comparison between lab-based and crowdsourcing subjective experiments.

The quality of the videos can also be compared against the average bitrate of the videos. This has been illustrated in Fig. 2, where the mean of the subjective scores has been calculated over the video contents. Generally, users prefer videos at higher bitrates, i.e., 3 or 5 Mbps and the difference between them is probably more due to the difference in content than the difference in compression levels. Users dislike buffering events and it seems that the frequency is more important than the total duration of these events (both videos at 3 Mbps with buffering have a total buffering time of 2s), which is in line with earlier studies e.g. [9]. But if the bitrate is high enough and the frequency of the buffering events is low enough, e.g., the 1F3M video, this seems to be a viable alternative to decreasing the bitrate of the video or having a constant low bitrate, e.g., 600 kbps or 1 Mbps.

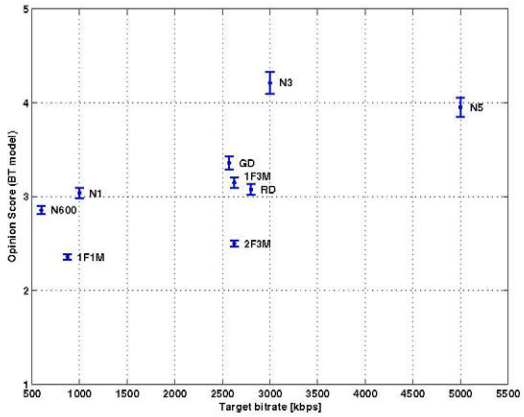

Fig. 2. Opinion Scores (BT model) versus the average bitrate.

4. CONCLUSION

The subjective experiment conducted in a crowdsourcing environment verifies the results of earlier studies of adaptation scenarios, including the effect of buffering events. Also, our study suggests that in a network environment with fluctuations in the bandwidth, a medium or low video bitrate which can be kept constant is the best approach. Moreover, if there are only a few drops in bandwidth, one can choose a medium or high bitrate with a single or few buffering events.

5. REFERENCES

[1] Cisco Visual Networking Index, "Global mobile data traffic forecast update, 2013-2018," *Cisco white paper*, 2014.

[2] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd - a framework for crowd-based quality evaluation," in *Picture Coding Symposium*, 2012, pp. 245–248.

[3] Tavakoli et al., "Subjective quality assessment of an adaptive video streaming model," in *IS&T/SPIE Electronic Imaging*. Int. Soc. for Optics and Photonics, 2014.

[4] "Rec. ITU-R BT. 500-13: Methodology for the subjective assessment of the quality of television pictures," 2012.

[5] "Rec. ITU-T P.910: Subjective video quality assessment methods for multimedia applications," 2008.

[6] J. Li, M. Barkowsky, and P. Le Callet, "Subjective assessment methodology for preference of experience in 3DTV," in *IVMSP Workshop*, 2013.

[7] Hossfeld et al., "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. on Multimedia*, vol. 16, no. 2, pp. 541–558, Feb 2014.

[8] R. Mok, E. Chan, and R. Chang, "Measuring the quality of experience of http video streaming," in *IFIP/IEEE Int. Symposium on Integrated Network Management (IM)*, 2011, pp. 485–492.

[9] S. Van Kester, T. Xiao, R. Kooij, K. Brunnström, and O. Ahmed, "Estimating the impact of single and multiple freezes on video quality," in *IS&T/SPIE Electronic Imaging*. Int. Soc. for Optics and Photonics, 2011.